Events Extracting From Stock Markets on the Web [*]

Dr. Eng. Yousef Abuzir [**]

Dr. Mohammad Dweib [***]

Dr. Eng. Yousef Sabbah [****]

Mr. AbdulRahman M. Baraka [*****]

[*] Received: 12/7/2018, Accepted: 4/10/2018.

[**] Professor/Al-Quds Open University/Palestine.
[***] Assistant /Al-Quds Open University/Palestine.
[****] Assistant /Al-Quds Open University/Palestine.
[*****] Instructor /Al-Quds Open University/Palestine.

DOI: https://doi.org/10.5281/zenodo.2582943

Abstract

In business and management, there is a need to understand the situations that occur in the stock markets domain to assist us in the analysis and investment decisions. This research focuses on financial markets and monitoring changes in stock prices by event recording, collecting, understanding their meanings, extracting information and organizing against its meanings within a well-organized structure. We have used four fields for each event: event name, entity name, attributes, and attribute values. Accordingly, we have proposed some approaches and rules for data analysis and evaluation to understand the events meanings in stock markets. Investors can depend on these events for decision making when buying or selling stocks. Based on the unstructured data we collected, we propose a set of rules and techniques to analyze, evaluate and understand the meaning of the events taking place in stock markets. These events can be used by the beneficiaries for decision making to buy or sell stocks. The results obtained are discussed and the performance is evaluated. Precision, recall and f-measure are used to evaluate our system, where the results of the conducted experiments show good performance of our proposed rules and techniques.

Keywords: Text Mining, Event Extraction, Stock Market, Prediction, Decision Making

استخراج الأحداث من سوق الأسهم
على الشبكة العنكبوتية

ملخص:

هناك حاجة في إدارة الأعمال لفهم المواقف التي تحدث في مجالات أسواق الأسهم كي يساعدنا ذلك في التحليل وقرارات الاستثمار. يركز هذا البحث على الأسواق المالية ومراقبة التغيرات الحادثة في أسعار الأسهم من خلال تسجيل الأحداث وجمعها وفهم معانيها، واستخراج المعلومات وتنظيمها بموجب هذه المعاني ضمن بنية حسنة التنظيم، وقد استخدمنا أربعة حقول لكل حدث هي: اسم الحدث، واسم ماهيته، وصفاته وقيم هذه الصفات. وبناءً على ذلك، قمنا باقتراح بعض المقاربات والقوانين لتحليل البيانات وتقييمها من أجل فهم معاني الأحداث في أسواق الأسهم، ويستطيع المستخدمون الاعتماد على هذه الأحداث في اتخاذ قراراتهم عند بيع الأسهم وشرائها. وبناء على البيانات غير المنظمة التي نجمعها، نقترح مجموعة من القوانين والأساليب لتحليل معاني الأحداث التي تجري في أسواق الأسهم وتقييم هذه المعاني وفهمها. ويمكن استخدام المستفيدين لهذه الأحداث في صنع القرار عند شراء الأسهم أو بيعها. وقد تمت مناقشة النتائج المستخلصة كما تم تحليل الأداء باستخدام الدقة والاسترجاع ومقياس–ف لتقييم نظامنا وأظهرت التجارب التي أجريت أداءً جيدًا لقواعدنا وتقنياتنا المقترحة.

الكلمات المفتاحية : التنقيب في البيانات، استخراج الأحداث، سوق الأسهم، التنبؤ، صنع القرار.

1 Introduction

The rapid increase in the use of the web and social networks is generating a huge amount of data that can be of a great value to different applications and users. However, it is becoming increasingly difficult for users to collect and analyze web pages that are relevant to a particular topic.

In the business and financial sector, data can be divided into three kinds: structured data, semi-structured data, and unstructured data. Unstructured text stores a lot of valuable financial information but lacks common structural frameworks. Usually, this text has many factors that increase the complexity of data processing and analysis like spelling errors, improper grammatical use and semantic ambiguities.

As shown in Figure 1 (Steven et al., 2009), the general process of information extraction (IE) includes text collection, text preprocessing, text manipulation and knowledge application. Text information extraction systems take the raw text of a document as its input and generates a list of (entity, relation, entity) tuples as its output.

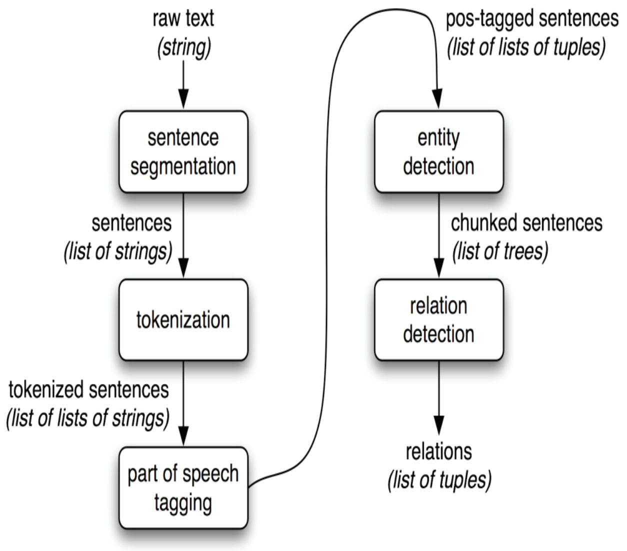


Figure 1: Main steps in Information Extraction

Our system is designed to acquire implicit knowledge that is hidden in the unstructured text. A wealth of valuable information can be discovered from web pages, such as stock markets investment or prediction about stock markets. The stock market is rapidly emerging as a sector that benefits considerably from event extraction using text mining techniques.

This paper proceeds as follows, section 2 is a literature review, while section 3 introduces the general idea about our problem and our systems. In section 4, we provide the research methodology. In section 5, we discuss the information extraction of stock markets based on text mining and research status of event recognition and extraction. In section 6, we introduce the main areas of the developed application and the results based on text mining. Finally, we summarize our model and analyze several aspects that need to be paid more attention and conclude the research in section 7.

2 Literature Review

Many researchers with different fields of specialization such as finance, computer science, artificial intelligence and other research communities (Roshni, Sagayam & Srinivasan, 2012), have focused on the kinds of information that can be useful for stock market prediction and the approaches in which this information can be retrieved.

There is a great interest and many advanced researches in the area of IE technology (Abuzir & Baraka, 2018) and (Abuzir, 2017). In the US, the Defense Advanced Research Projects Agency (DARPA) sponsored the Tipster Text Program and the Message Understanding Conferences (MUC). Both were the driving force behind the development of the technology (Gee, 1998). The MUC specifications for IE have become the actual standards in the IE research community. The MUC have divided IE into five distinct tasks: Named Entity (NE), Template Element (TE), Template Relation (TR), Co-reference (CO), and Scenario Templates (ST) (Chinchor & Marsh, 1998). Varieties of systems and techniques have been developed to address IE, where rule-based methods that employ some form of machine learning have been especially popular in the recent years (Wijnand, Viorel & Frederik, 2014).

Asilkan et al. have used RapidMiner for unstructured text mining and visualization, which includes finding correlation between the attributes of people in a survey dedicated to particular individuals. They have determined the most weighted attribute and used clustering to classify a group of the most similar people based on their attributes. They have also examined the above procedure by definitions, examples, analysis and conclusions (Asilkan, Ismaili & Nuredini, 2011).

Attia et al. have adapted and extended the automatic Multilingual, Interoperable Named Entity Lexicon approach to Arabic. For this purpose, they have used Arabic WordNet (AWN) and Arabic Wikipedia (AWK). AWN and AWK pass into five steps (Attia, Toral, Tounsi, Monachini & Genabith, 2010):

- ♦ Extraction of AWN’s instantiable nouns and identifying the corresponding categories and hyponym subcategories in AWK;
- ♦ Identifying named entities (NEs) through exploitation of Wikipedia links to find matches between articles in ten different languages;
- ♦ Applying keyword search on AWK abstracts for Arabic articles that do not have a match in other languages;
- ♦ Post-processing to obtain further NEs from AWK which is not reachable through AWN;
- ♦ Investigation of diacritization by matching

with geonames databases, MADA-TOKAN tools and different heuristics for restoring vowel marks of Arabic NEs.

Chibelushi and Thelwall (2009) have examined whether text-mining techniques can be used to extract decision-making elements from the meetings of software development project to help in developing a decision management system. They have presented and used theories of discourse, lexical chaining and cohesion, information retrieval and data mining methods for the analysis of meeting transcripts. Moreover, they have assessed the performance of the algorithm using TextTiling algorithm. Results of their method have shown that it is able to identify and extract decision-making needs and actions with a recall of 85–95% and a precision of 54-68%.

According to Hogenboom, Frasincar, Kaymak, and de Jong (2011), the main approaches to event extraction in the form of text mining are data-driven event extraction, knowledge-driven event extraction, and hybrid event extraction. With the advances of Natural Language Processing (NLP) techniques, various studies have found that financial news can dramatically affect the share price of a security (William & Zhenhao, 2014), (Ronny & Alexandre, 2012) and (Boyi, Rebecca, Leon & Germ´an, 2013) .

Text mining refers to the extraction of useful information and knowledge from unstructured text. Researchers discuss text mining as a field of information retrieval, machine learning, statistics, computational linguistics and data mining. They summarize the analysis tasks, which include preprocessing, classification, clustering, information extraction and visualization. Moreover, they discuss several applications of text mining. Finally, they refer to classification model performance measures as follows (Hotho, Nürnberger & Paaß, 2005):

- Precision (P): number of correctly extracted events divided by the total number of extractions.
- Recall (R): number of correctly extracted events divided by the total number of answer extractions.
- F measure: this measure depends on both P and R, which equals to 2PR/(P+R).

Bollen, Mao and Zeng (2011)have used a Granger causality analysis and a Self-Organizing Fuzzy Neural Network to investigate if public mood states derived from Twitter feeds are predictive of changes in Dow Jones Industrial Average (DJIA) closing values. They have employed two mood-tracking tools to analyze daily tweets: Opinion Finder to measure positive vs. negative moods and Google-Profile of Mood States (GPOMS) to measure moods. Their results have indicated that the inclusion of specific public mood dimensions improves the accuracy of DJIA predictions. The prediction accuracy achieved 87.6% for daily up and down changes in the closing values and reduced the Mean Average Percentage Error by 6%.

Jungermann (2011) has proposed and developed an IE extension for RapidMiner open source framework. He has presented how to install the plugin and use it with RapidMiner, which works with specific data structure of datasets. The plugin operates with the process of data mining tasks in four parts, data retrieval, pre-processing (i.e. data preparation), modeling of pre-processed data, and finally, evaluation of the models to select the optimal one.

Khandelwal et al. have focused on identifying stories with similar topic, which resulted in a very large number of reported stories. They have assumed it more useful to present a list of main events in the topic than the entire collection of stories. Moreover, they have proposed a scheme and developed a test-bed of user judgments to evaluate the output of their techniques. Finally, they have created a corpus that can also be used to evaluate single or multi-document summaries (Khandelwal, Gupta & Allan, 2001).

In another research, Jegadeesh and Titman (1993) have suggested that buying stocks which have performed well and selling stocks that have performed poorly in the past generate significant positive returns. Such strategies have proved that the profitability is related neither to systematic risk nor to delayed stock price reactions to common factors. Nevertheless, part of the weird returns in the first year scatters in the following two years. A similar sample of revenue returns has been also documented for former winners and losers.

Roshni, Sagayam & Srinivasan (2012) have presented a survey of text mining, which operates by converting unstructured words and phrases into numerical values. They have proposed to link these numerical values with structured data in a database and analyze the data using traditional data mining methods, which make text information retrieval and data mining dramatically important (Roshni, Sagayam & Srinivasan, 2012; Saravanan & Chonkanathan, 2010). In addition, Saravanan and Chonkanathan (2010)have presented the use of text mining with a novel high dimensional clustering algorithm that provides an exploratory data mining associated with the text. They have also introduced results of their experiments for analyzing a real-world text dataset.

Abuleil and Alsamara (2015) have targeted stock markets domain to observe, record, collect events, clarify their meanings, and organize the information in a well-structured format. They have used Semantic Role Labeling (SRL) to identify four factors for each event: verb of action, entity name, attribute and attribute value. Based on the structured data collected, they have proposed a set of rules and techniques to analyze, evaluate and understand the meaning of the events taking place in stock markets.

In his research, Soni (2011) has surveyed the previous studies to predict stock market changes based on machine learning and artificial intelligence techniques. He has suggested Artificial Neural Networks (ANNs) as the primary technique of machine learning to predict the raise and drop of the prices of the stocks in the financial market.

De Bondt and Thaler (1985) have investigated whether the behavior of people, who tend to “overreact” to unexpected events, affects stock prices. Based on the monthly return data of the Center for Research in Security Prices (CRSP), the empirical evidence is consistent with the overreaction hypothesis. Their results have shown the January returns earned by prior “winners” and “losers”, where losers’ portfolios were exceptionally large as the five years after portfolio formation.

Even though various technical, fundamental and statistical indicators have been proposed and used with varying results, none of the existing systems is dedicated to understand the situations that occur in the stock markets domain to assist us in the analysis and investment decisions. Moreover, none of the existing systems identified the four fields for each event in this study: verb of action, entity name, attribute, and attribute value, which proposes set of rules and techniques to analyze, evaluating and understand the meaning of the events that taking place in stock markets.

3. Problem Statement and Research Questions

The state of the art event processing approaches deal primarily with the syntactic processing of low-level primitive events, constructive event database views, streams and primitive actions. The identification of critical events requires processing of a huge amount of data and metadata. For some application scenarios, an intelligent event processing engine is required which can understand what happens in terms of events and can (process) state and know what reactions and processes it can invoke, and which new events it can signal or predict.

The detailed research questions related to this problem are addressed below:

- ♦ How should raw events and complex events (event patterns) in stock market on the web be represented on financial resources?
- ♦ How should knowledge about events and event patterns be represented?
- ♦ What can be an adequate representation of events, actions, states, situations and other related concepts for event extraction?

3.1 Research Importance

Prediction stock price or financial markets have been one of the biggest challenges to the investors’ community. Various technical, fundamental, and statistical indicators have been proposed and used with the varying results. Text mining is identified as one of the techniques used in the field of stock market prediction.

The main reason of using text mining is to try to help the investors in the stock market to decide the best timing for buying or selling stocks based on the knowledge extracted from the historical prices of such stocks. The taken decision will be based on one of the text mining techniques.

Forecasting financial stock market (Abuzir & Baraka, 2018), currency exchange rate, bank bankruptcies, understanding and managing financial risk, trading futures, credit rating, loan management, bank customer profiling, and money laundering analyses are core financial tasks for data and text mining. Most of these tasks many benefit or have similarities with text mining techniques in this research.

There are different advantages for using our system for the event extraction of real world unstructured financial news posted on the web. We can benefit from the research as follows:

- ♦ Our students especially, in finance and business can benefit from the system and can learn more about stock market investment.
- ♦ The private and public sectors like universities, banks, companies, schools and training centers can use our model to extract events and stock markets for performance forecasting
- ♦ The model can be used in practical parts of related courses.
- ♦ Disseminating our findings in different journals and conferences.
- ♦ Organizing workshops for dissemination of the results and findings of our research.

3.2 Research Objectives

The main objective of this research is to identify, investigate, analyze and evaluate valuable features and methodologies in stock market movement performance forecasting specific decision. This can be achieved through the following objectives:

- ♦ Study the existing stock market systems that are based on financial news articles published on the web with the focus on systems that use text-mining methods.
- ♦ Investigate text-mining methods that have been used in the development of our system.
- ♦ Design, implement and evaluate the proposed system that applies text mining on news articles.

We have conducted a case study and experiments based on text mining or using feature selection and web analysis. In context analysis, the performance of the proposed system is evaluated based on these experiments.

4. Methodology

In this research, the exploratory research and systematic literature review have been used to analyze and apply the results of our systems for event extraction from stock markets. Moreover, we have applied the various methods of text mining to study the performance of event extraction from the web related to stock markets to assist in analysis and prediction of investment decisions. We have developed a prototype of a text-mining tool using the proposed methods. We have also employed experimental research to evaluate these methods using standard scoring measures of information retrieval.

4.1 Research Instruments

Many tasks have previously performed to build text-mining models. In this research, we have used text-mining methods to build an event detector and analyzer. We have used Microsoft Visual C# to develop the proposed system and MySQL database engine to store the extracted events data.

4.2 Research Scope

As previously mentioned, the main objective of this research is to analyze the historical data of stock markets available on the web using text-mining methods, in order to help investors to know when to buy new stocks or to sell their stocks. Hence, this research targets the investors in the stock and financial markets as well as the interested researchers in this field. Stock market analysis requires large historical datasets to ensure reliability. In order to evaluate our text-mining model, we have used a sample from historical stock and financial data available on the related websites. Because of limited resources of specialized data and samples, the scope of this data has been restricted to particular financial markets and a specific period. We have focused on the financial sector and the companies in

MarketWatch News, forex and Yahoo Finance for the period (April 2016 - April 2018) on the following websites:

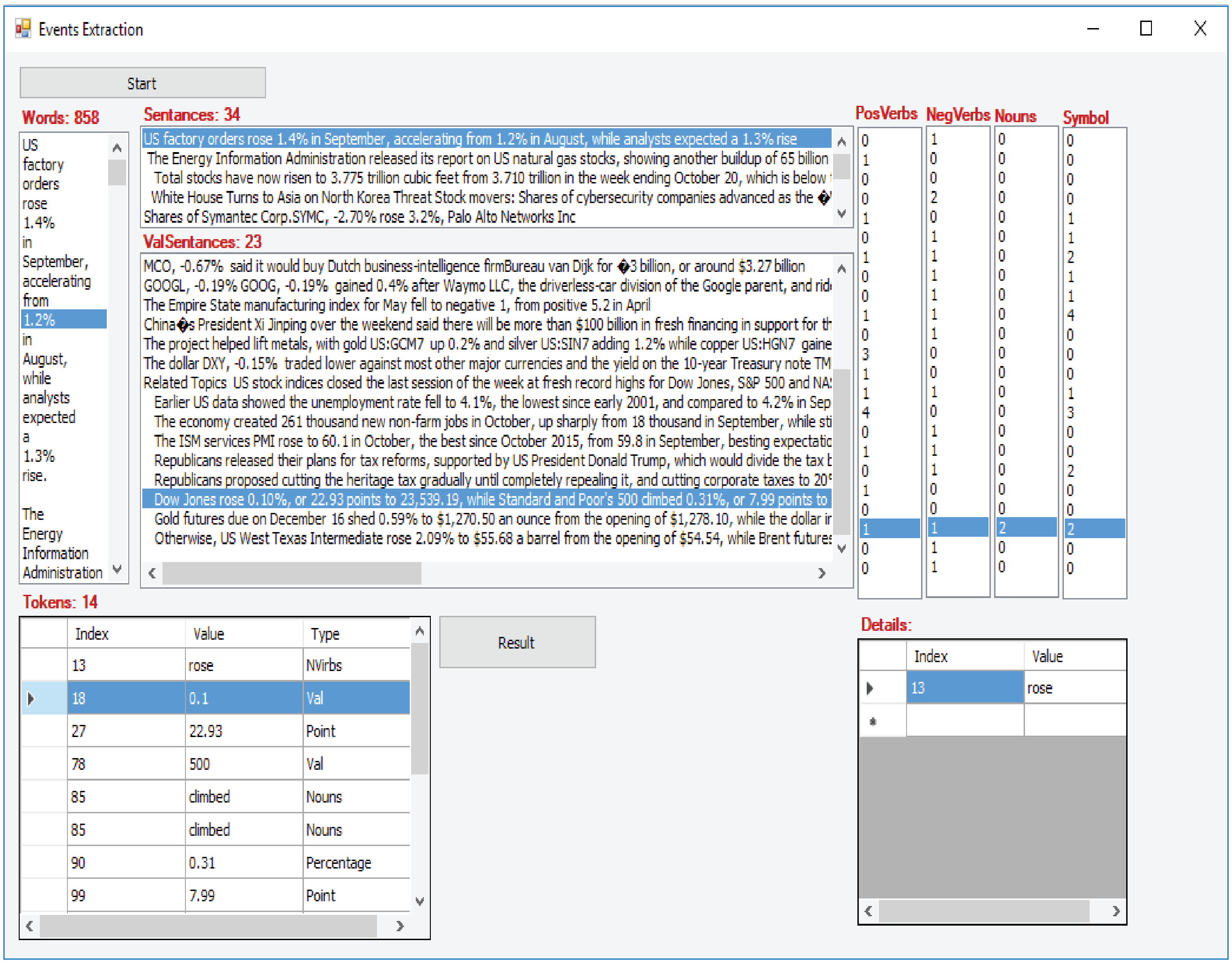
- marketwatch.com
- forex.com
- finance.yahoo.com

The datasets consist of daily opening, high, low and closing prices and have been adjusted for stock splits and dividends.

5 The Proposed System

Our proposed system i.e. Events Extraction System (EES) has been developed using Microsoft Visual C# and Microsoft Visual Studio.net IDE (Integrated Development Environment). In the first phase of this research, we have developed the main interface of the EES, which consists of nine windows, as shown in figure 2. Each window (i.e. list) has a specific title, which refers to its meaning. The number of windows may increase in next stages.

In the second phase, we have developed the data collection method and the proposed model. For data collection method, we have chosen some stock and financial markets websites and collected the example paragraphs. After that, we have developed our proposed model using Microsoft Visual C#, since RapidMiner software is complicated to solve all the cases, and in some cases, we encountered some sever troubles. ESS reads paragraphs, applies some prepossessing and identifies five types of tokens: Positive Events, Negative Events, Entities, Nouns and Numbers.



Words: 858

US
factory
orders
rose
1.4%
in
September,
accelerating
from
1.2%
in
August,
while
analysts
expected
a
1.3%
rise.

The
Energy
Information
Administration

Sentances: 34

US factory orders rose 1.4% in September, accelerating from 1.2% in August, while analysts expected a 1.3% rise
The Energy Information Administration released its report on US natural gas stocks, showing another buildup of 65 billion
Total stocks have now risen to 3.775 trillion cubic feet from 3.710 trillion in the week ending October 20, which is below
White House Turns to Asia on North Korea Threat Stock movers: Shares of cybersecurity companies advanced as the �
Shares of Symantec Corp.SYMC, -2.70% rose 3.2%, Palo Alto Networks Inc

ValSentances: 23

MCO, -0.67% said it would buy Dutch business-intelligence firmBureau van Dijk for �3 billion, or around $3.27 billion
GOOGL, -0.19% GOOG, -0.19% gained 0.4% after Waymo LLC, the driverless-car division of the Google parent, and rid
The Empire State manufacturing index for May fell to negative 1, from positive 5.2 in April
China�s President Xi Jinping over the weekend said there will be more than $100 billion in fresh financing in support for th
The project helped lift metals, with gold US:GCM7 up 0.2% and silver US:SIN7 adding 1.2% while copper US:HGN7 gaine
The dollar DXY, -0.15% traded lower against most other major currencies and the yield on the 10-year Treasury note TM
Related Topics US stock indices closed the last session of the week at fresh record highs for Dow Jones, S&P 500 and NA
Earlier US data showed the unemployment rate fell to 4.1%, the lowest since early 2001, and compared to 4.2% in Sep
The economy created 261 thousand new non-farm jobs in October, up sharply from 18 thousand in September, while sti
The ISM services PMI rose to 60.1 in October, the best since October 2015, from 59.8 in September, besting expectatio
Republicans released their plans for tax reforms, supported by US President Donald Trump, which would divide the tax b
Republicans proposed cutting the heritage tax gradually until completely repealing it, and cutting corporate taxes to 20
Dow Jones rose 0.10%, or 22.93 points to 23,539.19, while Standard and Poor's 500 climbed 0.31%, or 7.99 points to
Gold futures due on December 16 shed 0.59% to $1,270.50 an ounce from the opening of $1,278.10, while the dollar ir
Otherwise, US West Texas Intermediate rose 2.09% to $55.68 a barrel from the opening of $54.54, while Brent futures

PosVerbs	NegVerbs	Nouns	Symbol
0	1	0	0
1	0	0	0
0	0	0	0
0	2	0	0
1	0	0	1
0	1	0	1
1	1	0	2
0	1	0	1
0	1	0	1
1	1	0	4
0	1	0	0
3	0	0	0
1	0	0	0
1	1	0	1
4	0	0	3
0	1	0	0
1	1	0	0
0	1	0	2
1	0	0	0
0	0	0	0
1	1	2	2
0	1	0	0
0	1	0	0

Tokens: 14

Index	Value	Type
13	rose	NVirbs
18	0.1	Val
27	22.93	Point
78	500	Val
85	climbed	Nouns
85	climbed	Nouns
90	0.31	Percentage
99	7.99	Point

Details:

Index	Value
13	rose

Figure 2: Interface of Events Extraction System (EES).

5.1 System Operation:

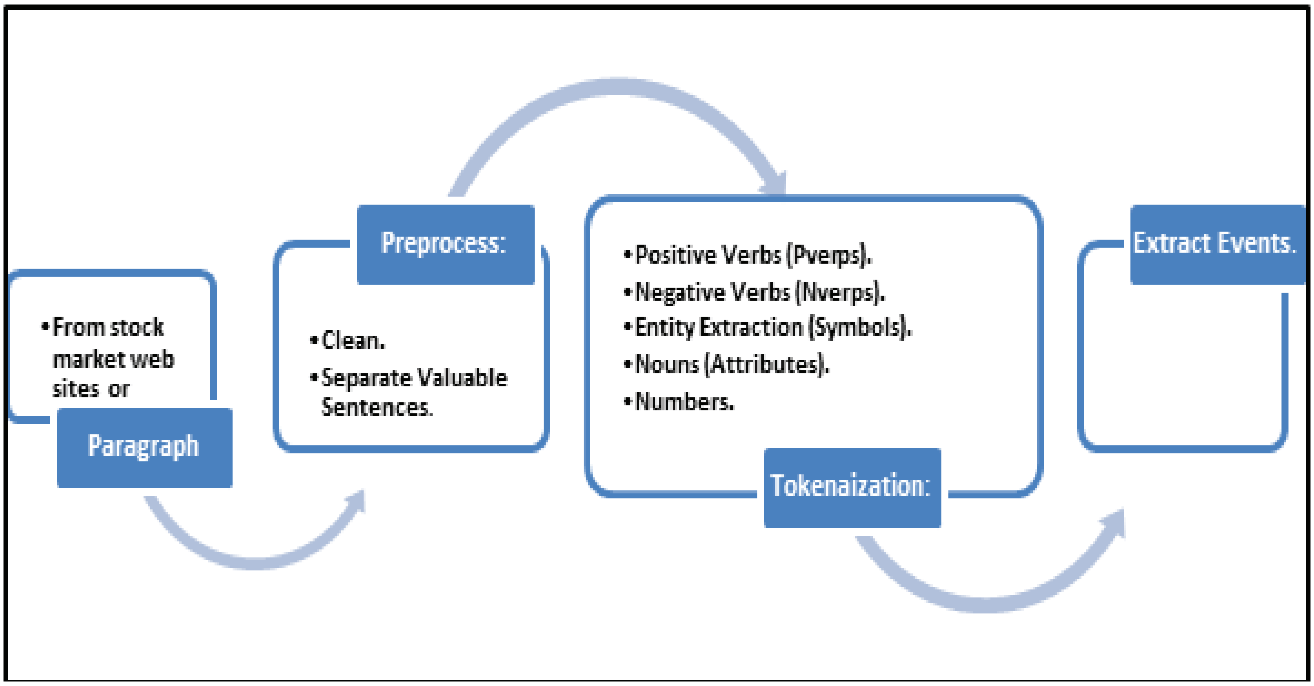
The system aims to identify and extract the required words (Tokens) throughout the paragraphs based on context analysis, as shown in figure 3. The analysis process passes into four phases as follows:

- Entry of a Paragraph: Selects an input stock market paragraph from the web or the dataset file.
- Pre-processing. It is the first task before term extraction from the paragraphs, which involves text cleaning and punctuation marks (e.g. hyphens and brackets) removal (Vijayarani & Janani, 2016). Accordingly, pre-processing passes through three steps, as follows:

    - **Paragraph cleaning:** At this stage, we remove all unwanted terms, letters or signs in a paragraph. This includes, for instance, spaces and hidden signs that appear when the files are translated into text variables, like "/r" and "/n". Also, we remove the comma "," that appears within numbers such as (2,510.00) to become (2510.00).

    - **Sentence splitting:** All sentences in each paragraph are split using the following punctuation marks that are usually used as delimiters between sentences:
        a. End of Line (EOL) sign.
        b. Dot (.).
        c. Comma (,).

    - **Exception handling:** All sentences in our system should contain numbers or values (e.g. currency, points or percentages), so we exclude the non-numeric sentences (i.e. sentences that do not contain any number or value).



•From stock market web sites or
Paragraph
Preprocess:
•Clean.
•Separate Valuable Sentences.
•Positive Verbs (Pverps).
•Negative Verbs (Nverps).
•Entity Extraction (Symbols).
•Nouns (Attributes).
•Numbers.
Tokenaization:
Extract Events.

Figure 3: EES Operational Phases

- Tokenization: This phase aims at breaking a stream of textual content up into words, terms, symbols, or some other meaningful elements called tokens (Vijayarani & Janani, 2016). This operation is trivial while the text is stored in machine-readable formats. Therefore, we have employed a parser (e.g. a tokenizer) to extract the words (e.g. tokens) in the text analysis. Tokens may be standardized using a dictionary to map different, but equivalent, variants of a term into a single canonical form (Witten, 2005). We have extracted many entities from the example shown in table 1, as follows:

    - **Positive Events:** Positive events refer to all words (i.e. events) that indicate increase in the trading volume. We have created a list of positive events as follows: "up", "rising", "highs", "increasing", "higher", "grew", "Rank in the top", "+", "high", "over", "add around", "more than", "strong", "stepped up", "won", "growth"," increase"," increased".

market, or a percentage of increment or decrement (%). We have identified numbers by two main factors: the measuring unit and the meaning.

- **Event Extraction:** A key element and a potential action of our proposed model is the integration of context analysis with semantic analysis. Semantic analysis extracts other semantic relations, which are related to stock markets analysis. Our system uses the Event Extraction model to discover these semantics in our dataset.

6 Results

The proposed system is the first to utilize text mining for addressing the need to understand the events that occur in the stock markets domain. It assists to observe and record changes (events) when they happen, collect them, understand their meanings, and organize the information along with meanings in a well-structured format. Moreover, it offers various core financial tasks using text mining, including: forecasting of stock market, currency exchange rate, bank bankruptcies, understanding and managing financial risks, trading futures, credit rating, loan management, bank customer profiling, and money laundering analysis. Based on processing a huge amount of data and metadata, the proposed system offers the following new features:

- Strong prediction model for event extraction in stock market, which can be utilized by an investor for better prediction when to sell or to buy.
- Advanced pattern discovery methods that deal with complex numeric and non-numeric data, structured objects, as well as text and data in a variety of discrete and continuous scales (nominal, order, absolute, raise, drop, and so on).
- The results prove the benefits of using such methods for stock market forecasting.
- A new trend in developing practical decision support systems that makes it easier to rely on data mining environment specified for financial tasks.
- The model could be embedded into other financial, banking, and investing systems.

- **Negative Events:** Negative events refer to all words (i.e. events) that indicate decrease in the trading volume. We have also created a list of negative events as follows: “down”, “fell”, “-”, “dropped to”, “dropped nearly”, “lost”, “less than”, “dropped”, ““drop”, “rose”, “arose”, “flat”, “retreated”, “declined”, “decrease”,” decreasing”.

- **Entities Extraction:** The entities or symbols are the core of events such as company name, bank name and stock market name. We have two types of Entities: The first type is any words added in list of symbols, the default list consists of: “COMP”, “DJIA”, “DJT”, “DJU”, “GDOW”, “IXCO”, “MID”, “NDX”, “NYA”, “OEX”, “PSE”, “RUI”, “RUT”, “SOX”, “SPX”, “TNX”, “TYX”, “XAU”, “XAX”, “XMI”, “XOI”, “DJIA”, “S&P”, “NASDAQ”, “CXP”, “Microsoft”, “Google”, “Amazon”. The second type is to add any word whose characters are all capital letters. We have collected a number of keywords and used them to recognize and tag the entities in the events.

- **Nouns or Attributes Extraction:** Attributes represent the main source of information about entities in the events. We have identified nine different types of attributes to describe the behaviour of events. The list of attributes includes: “percentage “, “price”, “worth”, “point”, “price-decrement”, “price-increment”, “climbed”, “percentage-decrement”, “percentage-increment”, “worth-decrement”.

- **Numbers or Values:** These are numeric values (i.e. numbers) of amount of currency or percentage, whether increasing or decreasing. Attribute values, which are mentioned in the events represent either a price (currency), a worth (number of points) which is the value of a certain entity in the stocks

We have evaluated our model on unstructured datasets extracted form online websites related to the domain of the research. We have employed standard measures commonly used in the research field of information retrieval to evaluate our model. These measures are: Precision, Recall and F measure.

6.1 Data Collection and Preparation

Often, a researcher in this domain builds his own dataset to apply and evaluate his model, because each researcher is interested in different stock market and companies and focuses on them. Although, we gathered samples to form our dataset from some stock and financial market websites [11-14]. The dataset consists of statements with different structures and contains many types of entities and variables. We have gathered 30 example files; each consists of a few paragraphs.

6.2 Experiments

We have proposed a system for automatic detection of changes that happen in the online stock markets domain, where phrases and subordinate information were selected that yield a high performance. The system is relatively simple and able to observe, detect and extract events from open domain unstructured text (Hotho, Nürnberger & Paaß, 2005). So far, this research has focused on stock markets domain expressed with unstructured sentences. The proposed system identifies four fields for each event, verb of action, entity name, attribute, and attribute value.

Different patterns have been used to extract information. Since the frequent patterns or features were built from the training set, we required to find a way for mapping of those patterns onto the test set of the unstructured text extracted form online websites. In fact, this is the crucial part of our system. Specifically, we needed to test our extracted patterns on a real sample of datasets, where each extracted sentence must map to a unique pattern. To accomplish that, we have developed an algorithm based on figure 3.

Table 1 illustrates the examples that have been tested using our proposed model and algorithm.

Table 1:

Sample of data set used in testing our model

Brent price dropped in October delivery and settled at 100.82 dollars a barrel.

The industrial index Dow Jones arose 72.10 points, by 6.0 % to close at 17078 points.

Following Monday's open, the Dow jumped 83.61 points, or 0.47% at 17,888.41; the S&P 500 Index gained 0.91 points or 0.04% at 2,071.93. The Nasdaq Composite added 6.39 points, or 7.58% to 4,773.69.

Gasoline prices dropped to the lowest since 2009, tumbling 24.68 cents in the two weeks ended Dec. 19 to $2.47 a gallon. The slide in oil prices continued on Thursday with Brent crude prices dropping below $90 a barrel for the first time in two years and West Texas...

As the price of Brent crude fell below $60 for the first time since 2009.

The first step in the process is pre-processing. After selecting a paragraph file and pressing the Start button, the system automatically separates the paragraph into numeric sentences. Accordingly, it depicts the numeric sentences in the ValSentences window and the number of sentences in the label of the ValSentences window. After that, Tokenization starts for all extracted sentences, which orders the extracted tokens in four windows, as follows:

- ♦ PosEvent: Window: number of Positive Event.
- ♦ NegEvent: Window: number of Negative Event.
- ♦ Symbols Window: number of Entity Extraction.
- ♦ Attributes (Nouns) Window: number of Nouns.

When any sentence in ValSentences list is selected, all tokens are extracted and listed in the Tokens Window, which shows two values: index of a token and its value or name.

A. Index of token: refer to position (index) of Token in sentence.

B. Value: refer to Token name or value (example: Up, "+" or 2500)

C. Type: indicate to Token Type. There are five types that were listed previously: Pvirbs, Nvirbs, Symbol, Nouns and Numbers. The Numbers type consist of three types: Percent, Pints and Val.

These attributes list is sorted by index (first column).

The Result button in our system saves all results to SQL Database, and Final_Table saves the original sentence in Org_Sentence column and other columns for rest results (Noun, Verb, Number and Noun) which are used for semantic meaning.

To evaluate our approach using our EES, 35 paragraphs were automatically parsed using the different phases for extraction of the different types of information. Then, we evaluated our approach using the basic evaluation measures of information retrieval systems, which can be calculated using equations (1-3).

$$\mathrm{Re}\,call = \frac{RR}{(RR + RNR)} * 100\,\% \qquad (1)$$

$$\mathrm{Pr}\,ecision = \frac{RR}{(RR + IRR)} * 100\% \qquad (2)$$

$$FMeasure = \frac{(\mathrm{Pr}\,ecision * \mathrm{Re}\,call)}{(\mathrm{Pr}\,ecision + \mathrm{Re}\,call)} \qquad (3)$$

Where RR is number of relevant events, RNR is number of relevant events not retrieved, and IRR is number of irrelevant events retrieved.

Table 3 shows the results of event extraction from our dataset. The “Actual relevant” field shows the actual number of events in our collection related to each type, as follows:

- Events.
- Type of numbers (e.g. a year, a price or a stock exchange).
- Percentage processing (e.g. if a number is a percentage, a stock exchange or a currency).
- Semantic analysis.

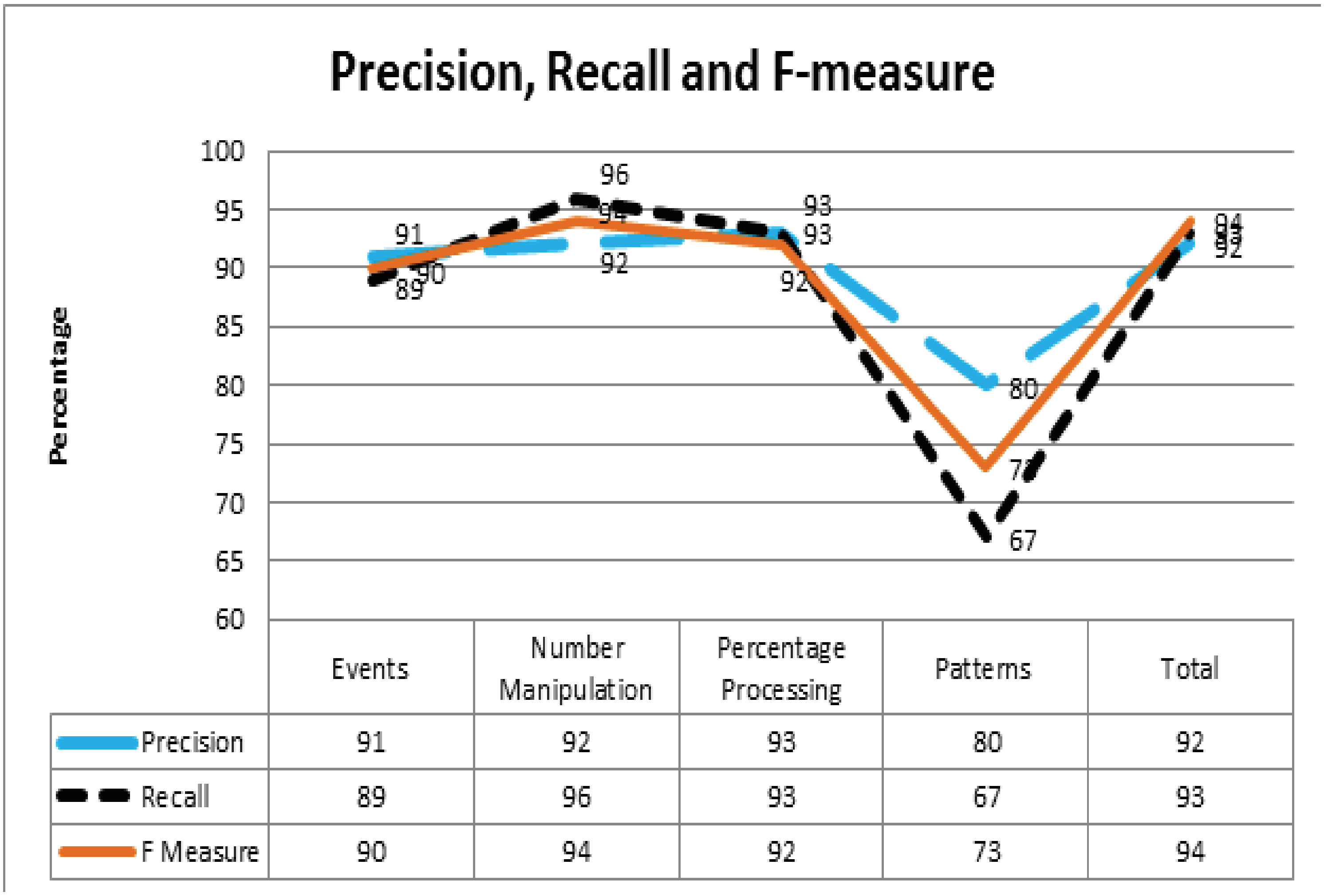
The second field shows the number of the retrieved relevant events of stock markets by the EES. The third field provides the number of actual events that are relevant to the stock markets events from the retrieved events. The last two fields show the number of relevant events that are not retrieved and the number of irrelevant retrieved events.

Table 2 shows the evaluation results based on recall, precision, and F-measure for each event type, and Figure 4 represents these results by graphs. This evaluation shows that our system has achieved acceptable results.

Table 2:

Evaluation of our system using precision, recall and F-measure

Type	Actual relevant	Number of retrieved	Relevant retrieved RR	Relevant not retrieved RNR	Irrelevant retrieved IRR	Precision	Recall	F Measure
						RR/ (RR+IRR)	RR/ (RR+RNR)	2*(Precision * Recall) / (Precision +Recall)
Events	36	35	32	4	3	0.914285714	0.888888889	0.901408451
Number Manipulation	134	140	129	5	11	0.921428571	0.962686567	0.941605839
Percentage Processing	60	55	51	5	4	0.927272727	0.910714286	0.918918919
Patterns	6	5	4	2	1	0.8	0.666666667	0.727272727
Total	236	235	216	16	19	0.919148936	0.931034483	0.925053533



	Events	Number Manipulation	Percentage Processing	Patterns	Total
Precision	91	92	93	80	92
Recall	89	96	93	67	93
F Measure	90	94	92	73	94

Figure 4: A graph representation of the evaluation results of our system using precision, recall and F-measure

7 Conclusions

This paper outlines a new approach for extracting events in unstructured text related to stock markets posted in online webpages. The approach is based on using predefined patterns as features to a map test data set to these patterns. Our approach has the advantage that features and textual structure are automatically discovered rather than manually selected. Nevertheless, it has the consequence that adding a new feature to this domain by reducing the manual effort requires exploring and extracting various combinations of events in stock markets domain. Our approach uses text mining techniques to extract events, capture the dependencies between features, and show the semantic behind them.

The obtained results are difficult to compare with other studies, since we focus on different patterns and fields. We have conducted an experiment using our proposed model and algorithm based on text mining techniques. In this experiment, we examined them on a real dataset with our set of patterns or features. In summary, the proposed model and algorithm have achieved good results. Moreover, we have evaluated the model based on recall, precision and F measure, which have also proved an acceptable performance.

References

1. Abuzir, Y. and Abuzir, S. (2017). Text Mining For Efficient Project Evaluation, International Conference on Education in Mathematics, Science & Technology (ICEMST), May 18 - 21, 2017 Ephesus-Kusadasi/Turkey.

2. Abuzir Y. and Baraka A. (2018). Financial Stock Market Forecast Using Data Mining in Palestine, accepted in Palestinian Journal of Technology and Applied Sciences, , No 2 (2018).

3. Asilkan, Ö., Ismaili, A. and Nuredini, K. (2011). An Exemplary Survey Implementation on Text Mining with Rapid Miner. 1st International Symposium on Computing in Informatics and Mathematics (ISCIM 2011), pp. 221-234.

4. Attia, M., Toral, A., Tounsi, L., Monachini, M. and Genabith J. (2010). "An Automatically Built Named Entity Lexicon for Arabic" LREC 2010 proceedings, Valletta, Malta, May 19-21, 2010.

5. Xie, B., Passonneau, R., Wu, L. and Creamer, G. (2013). Semantic frames to predict stock price movement. In Proc. of ACL, pages 873–883, 2013.

6. Chibelushi, C. and Thelwall, M. (2009). Text Mining for Meeting Transcript Analysis to Extract Key Decision Elements. Proceedings of the International Multi-Conference of Engineers and Computer Scientists 2009, Vol I, pp.710-715.

7. Chinchor, N. and Marsh, E. (1998) MUC-7 Information Extraction Task Definition (version 5.1), In Proceedings of MUC-7.

8. Hogenboom, F., Frasincar, F., Kaymak, U. and de Jong, F. (2011). "An Overview of Event Extraction from Text" 10th International Semantic Web Conference, Bonn, Germany. 2011.

9. Gee, F. (1998, October). The TIPSTER text program overview. In Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998 (pp. 3-5). Association for Computational Linguistics.

10. Hotho, A., Nürnberger, A. and Paaß, G. (2005). A Brief Survey of Text Mining. Journal for Computational Linguistics and Language Technology 20, pp. 19-62.

11. http://www.economies.com

12. http://www.ibtimes.com/economy

13. http://www.ibtimes.com/wall-streets-week-ahead-dow-jones-industrial-average-soars-santa-claus-rally-2014-1764746

14. http://www.marketwatch.com

15. Bollen, J., Mao, H. and Zeng, Z. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1):1–8.

16. Jungermann, F. (2011). Documentation of the Information Extraction Plugin for RapidMiner [online]. [Accessed 30 March 2014]. Available at: http://www-ai.cs.uni-dortmund.de/PublicPublicationFiles/jungermann_2011c.pdf

17. Khandelwal, V., Gupta, R. and Allan., J. (2001). An evaluation corpus for temporal summarization. In James Allan, editor, Proceedings of HLT 2001, First International Conference on Human Language Technology Research, San Francisco, 2001, Morgan Kaufmann.

18. Jegadeesh, N. and Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. The Journal of Finance, 48(1):65–91.

19. Luss, R. and d'Aspremont, A. (2012). Predicting abnormal returns from news using text classification. Quantitative Finance, (doi:10.1080/14697688.2012.672762):1–14, 2012.

20. Roshni, S., Sagayam R., and Srinivasan, S. (2012). A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques. International Journal of Computational Engineering Research, Vol. 2 Issue. 5 pp. 1443-1446.

21. Vijayarani S. and Janani, R. (2016). "Text Mining: Open Source Tokenization Tools – An Analysis". Advanced Computational Intelligence: An International Journal (Acii), Vol.3, No.1 (2016).

22. Abuleil, S. and Alsamara, K. (2015). "Collect Meaningful Information about Stock Markets from the Web" Journal of Systemics, Cybernetics and Informatics, 13(1)84-90, 2015.

23. Saravanan, D. and Chonkanathan, K. (2010). Text Data Mining: Clustering Approach. International Journal of Power Control Signal and Computation (IJPCSC), Vol. 1 No. 4, pp.

13-16.

24. Soni, S. (2011). Applications of ANNs in Stock Market Prediction: A Survey, International Journal of Computer Science & Engineering Technology (IJCSET), pp 71-83, Vol. 2 No. 3, 2011.

25. Bird, S., Klein, E. and Loper, E. (2009). Natural Language Processing with Python, O'Reilly 2009.

26. Bondt, W. and Thaler. R. (1985). Does the stock market overreact? The Journal of finance, 40(3):793–805.

27. Nuij, W. Milea, V. and Hogenboom, F. (2014). "An Automated Framework for Incorporating News into Stock Trading Strategies" IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 4, April 2014.

28. Yang Wang, W. and Hua, Z. (2014). A semiparametric gaussian copula regression model for predicting financial risks from earnings calls. In Proc. of ACL, pages 1155–1165, 2014.

29. Witten, I. (2005). "Text mining." in Practical handbook of internet computing, edited by M.P. Singh, pp. 14-1 - 14-22.(2005).